Actlets: A novel local representation for human action recognition in video

Muhammad M. Ullah, Ivan Laptev

ACTLETS: A NOVEL LOCAL REPRESENTATION FOR HUMAN ACTION RECOGNITION IN VIDEO

Muhammad Muneeb Ullah and Ivan Laptev

INRIA - Willow Project, Laboratoire d'Informatique, École Normale Supérieure, France

ABSTRACT

This paper addresses the problem of human action recognition in realistic videos. We follow the recently successful local approaches and represent videos by means of local motion descriptors. To overcome the huge variability of human actions in motion and appearance, we propose a supervised approach to learn local motion descriptors – *actlets* – from a large pool of annotated video data. The main motivation behind our method is to construct action-characteristic representations of body joints undergoing specific motion patterns while learning invariance with respect to changes in camera views, lighting, human clothing, and other factors. We avoid the prohibitive cost of manual supervision and show how to learn actlets automatically from synthetic videos of avatars driven by the motion-capture data. We evaluate our method and show its significant improvement as well as its complementarity to existing techniques on the challenging UCF-sports and YouTube-actions datasets.

Index Terms— Action recognition, local motion descriptors, supervised learning, actlets.

1. INTRODUCTION

Recognition of human actions in video is a highly challenging problem due to the large variations in person appearance, camera view points, backgrounds and other factors. An explicit way to address this problem is to reason about locations, poses, motion and interactions of people. While such a structural approach is appealing, it is difficult to implement in practice due to the modest performance of current methods for person detection and pose estimation in realistic video.

One alternative consists of representing human actions using statistics of local video descriptors. Despite being direct and simple, Bag-of-Features (BoF) type methods have recently been shown successful when applied to action recognition in realistic and challenging video data [1, 2, 3]. While the basic BoF model represents actions by disordered collections of local video descriptors, several ideas have been proposed aiming to improve BoF by modeling the spatial and temporal structure of activities [4, 5]. Complementary to these methods, the goal of this paper is to design improved local video descriptors providing better building blocks for statistical action representations.

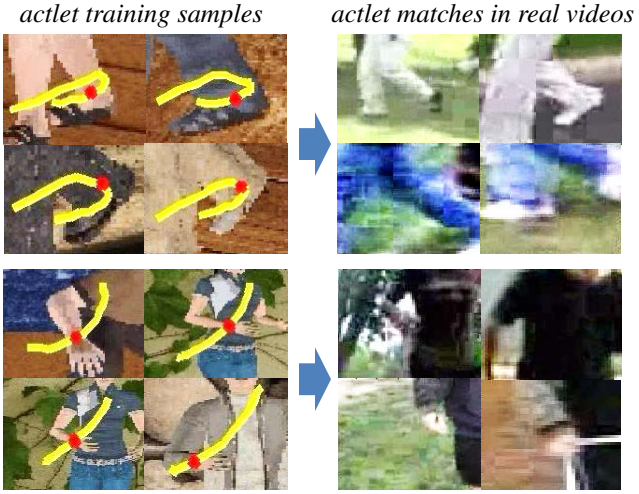

Fig. 1. Illustration of *actlets*, trained on synthetic data (left) and localized on the real videos (right). The automatically annotated trajectories of body-joints are shown on the left.

Local video representations, e.g., [6] are typically designed using unsupervised k-means quantization of local video descriptors into visual vocabularies [7] based on visual similarity. The rationale behind visual vocabularies is to group similar local events, such as upwards movements of hands, and to score occurrences of such events in subsequent recognition. While visual descriptors, e.g., HOG/HOF [1, 3] provide some invariance to variations of events in motion and appearance, unsupervised clustering may not be able to group similar events given the frequent variations of the video data due to changes in view points, lighting, background, clothing of people, style of actions and other factors.

Bourdev et al. [8] have recently proposed a supervised approach to learn appearance of body parts in static images. Body part detectors called *poselets* are trained to be invariant to irrelevant appearance variations using manual annotation of body parts in training images. Inspired by this representation, we in this paper propose a supervised approach to learn *actlets*, i.e., detectors of body parts undergoing specific patterns of motion. Learning actlets requires a substantial amount of annotated training data. To collect such data,

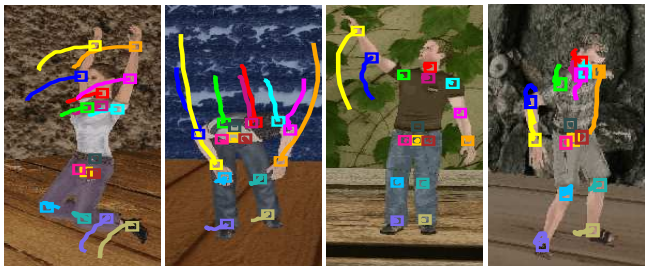

Fig. 2. Sample frames from our synthetic dataset illustrating variability of generated videos in terms of view points, backgrounds, character clothing and motion. Color curves illustrate automatically annotated trajectories of body joints.

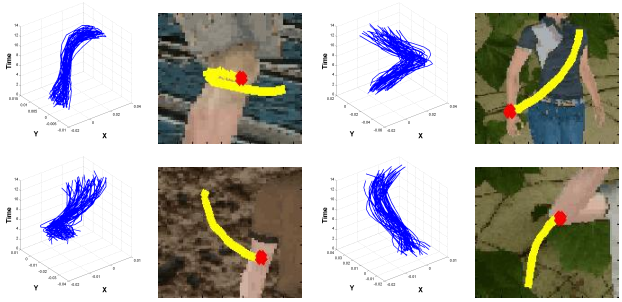

Fig. 3. Illustration of motion clusters for four body joints. All trajectories within a cluster are shown in separate plots by blue curves. An example patch for each cluster is also shown.

we here propose to avoid the heavy burden of manual video annotation and generate annotated data automatically by synthesizing videos of avatars driven by the motion-capture data (cf. Fig. 1). We next develop a method that successfully deploys actlets for action recognition. We evaluate our method and demonstrate its significant improvement as well as complementarity to existing techniques on the UCF-sports and YouTube-actions datasets.

The rest of the paper is organized as follows: Sec. 2 presents details of our synthetic dataset used to train actlets in Sec. 3. Sec. 4 describes application of actlets to action recognition. Sec. 5 presents evaluation of our method. Sec. 6 concludes this paper with a discussion.

2. SYNTHETIC DATASET OF HUMAN MOTIONS

To train a representative set of actlets, we need a relatively large amount of training data. The training data should cover a diverse range of human movements and should contain annotated positions of body joints over time. Also, a significant amount of variation in terms of appearance (e.g. clothing and background), view-point, illumination, and camera motion, is required to span the expected variability of the test videos. While manual annotation of body joints and their motion in video is highly time-consuming and therefore impractical, we resort to animation techniques and use motion capture data to build a synthetic dataset. The main advantage in this approach, is the availability of the ground-truth positions of body-joints in each synthesized video provided by the 2D projections of 3D body-joint positions of the motion-capture data. We use the CMU motion capture database[1], containing a large number of human motion sequences; from simple locomotions and physical activities to more complex movements involving human interactions.

We perform motion re-targeting of CMU motion capture sequences on 3D humanoid characters in Autodesk MotionBuilder 2011, and render videos from a set of fixed locations. We use ten 3D characters including males and females of different physiques, wearing different clothes. We render videos from a set of three different camera view points (front, left and right with respect to the character) while using five different static backgrounds. Additionally, we simulate the panning of the camera which follows the motion of the character in each video. We render one video for each motion capture sequence in the CMU database while randomly choosing a character, background, and a view point; and get 2549 synthetic video sequences in total. Fig. 2 illustrates a few example frames from our synthetic dataset together with the automatically annotated trajectories of body joints. We will make our dataset publically available upon acceptance of this paper.

[1]Available at: `http://mocap.cs.cmu.edu`

3. TRAINING ACTLETS

In this paper we consider the motion of nine body-joints (head, left/right elbow, left/right wrist, left/right knee and left/right ankle), as these are expected to provide rich action description. To group body joints with similar motion patterns, we perform clustering of 2D trajectories associated with each of the nine body joints. We then extract video patches for each trajectory and use them to train one actlet classifier for each trajectory cluster. The details of the method are described below.

3.1. Trajectory representation

For each of the nine body-joints in a synthetic video, the associated 2D trajectory with spatial coordinates (x_t, y_t) over time $t \in 1...T$ is subdivided into overlapping sub-trajectories, each having a length of $L = 15$ frames. The shape of a sub-trajectory encodes the local motion pattern associated with the body-joint. Following [9], we represent the shape of a sub-trajectory with a velocity-based vector. Given a sub-trajectory of length L, we describe its shape by a sequence $S = (\Delta P_t, \ldots, \Delta P_{t+L-1})$ of displacement vectors $\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$. The resulting vector S is normalized by the height of the character in the rendered video.

3.2. Clustering and training of actlets

We perform k-means clustering (we set $k = 75$) on all sub-trajectories associated with each of the nine body joints in all 2549 synthetic videos. We avoid occluded body joints and remove trajectories of right/left joints from the videos synthesized for the left/right views of the person respectively. We perform both view-specific and view independent clustering, where trajectories from the three different views are clustered either separately or jointly. To select distinct clusters, we sort clusters for each body joint according to the decreasing sum of distances to other clusters and keep the top $n = 50$ clusters from them. Fig. 3 illustrates examples of our clusters.

To train an actlet for a given body joint and motion pattern, we extract video patches in the neghbourhood of trajectories from one trajectory cluster. These video patches serve as positive training samples for an actlet. For the negative training samples, we randomly extract 10,000 synthetic video patches, corresponding to trajectories from the remaining 49 clusters of the same body joint. We represent extracted video patches by histograms of optical flow (HOF) descriptors [1].[2] We then train a linear SVM classifier for the HOF descriptors. This way, we obtain a total of 1000 linear SVM classifiers[3], corresponding to the view-specific and view-independent actlets.

4. CLASSIFICATION

Actlets provide a means to detect specific motion patterns of body joints in video disregarding irrelevant variations of the data in terms of backgrounds, clothing, view points and other factors. Our next goal is to deploy such descriptors for action recognition in real video. Given a video, we extract densely-sampled video patches and represent them by the HOF descriptors. For each HOF descriptor we obtain a set of actlet scores according to all trained actlet classifiers. We then use a 24-level spatio-temporal grid and concatenate maximum response value of each actlet within each grid cell into a vector representing the whole video. We refer to this video representation the *Actlet channel*. A somewhat similar approach called object banks has been previously proposed to represent still images [15]. For action classification based on the Actlet channel, we use a non-linear SVM [13] with RBF kernel.

We use Bag-of-Features (BoF) video representation as a baseline. BoF is typically based on k-means clustering, which is used to quantize local spatio-temporal descriptors into visual vocabularies, based on visual similarity. Here, we follow [11], and build the Bag-of-Features video representation using the Harris3D detector [6] in combination with the HOG/HOF descriptors [1]. We refer to this video representation as *BoF channel* (see [4] for more details). For classification, we use a non-linear SVM [13] with a χ^2 kernel [1]. We use one-against-rest approach for multi-class classification for the both channels.

[2]We use motion descriptors only as we expect motion, in contrast to appearance, to transfer well between synthetic and real videos.

[3]Front: 9 joints$\times$50 clusters + left/right: 2$\times$5 joints$\times$50 clusters + view-independent: 9 joints$\times$50 clusters. We train actlets for clusters with the minimum of 50 trajectories.

Fig. 4. Sample frames from video sequences of UCF-Sports (top), and YouTube Actions (bottom) datasets.

We integrate the Actlet channel with the BoF channel using multi-channel kernel[14]:

$$K(x_i, x_j) = exp\left(-\sum_c \frac{1}{\Omega_c} D\left(x_i^c, x_j^c\right)\right) \quad , \qquad (1)$$

where $D(x_i^c, x_j^c)$ is the distance computed using video channel x^c between videos i and j, and Ω_c is the normalization factor computed as an average channel distance [14]. In our case, $D(x_i^c, x_j^c)$ is the Euclidean distance for the Actlet channel, whereas, χ^2 distance for the BoF channel.

5. EXPERIMENTS

We evaluate the performance of actlets on the task of action classification in two challenging datasets: UCF-Sports and YouTube Actions. The two datasets mainly contain sports action classes.

The **UCF-Sports** dataset [10] contains 10 different types of human actions: swinging (on the pommel horse and on the floor), diving, kicking (a ball), weight-lifting, horse-riding, running, skateboarding, swinging (at the high bar), golf swinging and walking (cf. Fig. 4, top). The dataset consists of 150 video samples which show a large intra-class variability. To increase the amount of data samples, we extend the dataset by adding a horizontally flipped version of each sequence to the dataset (similar to [11]). We train a multi-class classifier and report the average accuracy over all classes. We use a leave-one-out setup and test on each original sequence while training on all other sequences together with their flipped versions (the flipped version of the tested sequence is removed from the training set).

The **YouTube Actions** dataset [12] contains 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking and walking with a dog (cf. Fig. 4, bottom). This dataset is challenging due to the large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background and illumination conditions. The dataset contains a

	BoF [%]	Actlet [%]	Comb. [%]
Dive	**100.00**	**100.00**	**100.00**
Golf swing	61.25	51.25	**90.00**
Kick ball	90.00	**100.00**	**100.00**
Weight lift	**100.00**	83.33	**100.00**
Horse ride	**75.00**	50.00	58.33
Run	**76.19**	52.38	69.05
Swing:pommel	85.00	**100.00**	95.00
Skateboard	0.00	**33.33**	25.00
Walk	90.91	81.82	**95.45**
Swing:high bar	**91.67**	**91.67**	**91.67**
Average accuracy	77.00	74.38	**82.45**

Table 1. Accuracy for the UCF-Sports dataset.

	BoF [%]	Actlet [%]	Comb. [%]
Bike	72.29	68.98	**83.87**
Dive	90.00	88.00	**92.00**
Golf	77.00	77.00	**87.00**
Soccer juggle	46.00	55.00	**59.00**
Trampoline jump	66.00	63.00	**73.00**
Horse ride	67.00	70.00	**73.00**
Basketball shoot	22.33	**42.00**	38.67
Volleyball spike	67.00	**81.00**	**81.00**
Swing	73.00	**79.00**	76.00
Tennis swing	51.00	44.00	**55.00**
Walk	34.83	39.90	**50.36**
Average accuracy	60.59	64.35	**69.90**

Table 2. Accuracy for the YouTube Actions dataset.

total of 1168 sequences. We follow the original setup [12] and perform leave-one-out cross validation for a pre-defined set of 25 folds. As for the UCF-Sports dataset, we report the average accuracy over all classes as the performance measure.

5.1. Results

Results on the UCF-Sports and YouTube datasets are presented in Table 1 and Table 2 respectively. We can notice that the performance of the Actlet channel is slightly lower than that of the baseline BoF channel on the UCF-Sports dataset, whereas, the Actlet channel performs better than the BoF channel on the YouTube dataset. Moreover, the combination of the Actlet channel with the BoF channel gives an improvement of approximately 6% over the BoF baseline in the case of UCF-Sports dataset and about 9% for the YouTube dataset. The better performance of the combined channels indicates their complementarity. Actlets focus on the characteristic local movements of people, whereas, BoF has a potential of capturing additional contextual information from the background. For the UCF-Sports dataset, we can observe that our proposed Actlet channel helped to significantly improve 5 out of 10 action classes (cf. Table 1), notably, the *Golf swing* and *Skateboard* action classes. On the YouTube dataset, the Actlet channel improved all the 11 action classes (cf. Table 2), specifically, the *Volleyball spike* class.

6. CONCLUSIONS

We have proposed a novel approach to represent local patterns of human motion in video, i.e., actlets. To get the required relatively large amount of annotated training data, we have avoided the expensive manual annotation and proposed to use synthetically generated videos of avatars driven by the motion capture data. We have shown how to train actlets for body parts undergoing particular motion, while making actlets insensitive to irrelevant variations in the video data. We have then proposed a new video representation based on actlets, and demonstrated significant improvement in human action recognition on the two challenging datasets, i.e., the UCF-sports and YouTube-actions datasets.

7. REFERENCES

[1] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.

[2] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in *ICCV*, 2009.

[3] H. Wang, A. Klaser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," in *CVPR*, 2011.

[4] M. M. Ullah, S. N. Parizi, and I. Laptev, "Improving bag-of-features action recognition with non-local cues," in *BMVC*, 2010.

[5] J.C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *ECCV*, 2010.

[6] I. Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2/3, pp. 107–123, 2005.

[7] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.

[8] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *ICCV*, 2009.

[9] R. Sukthankar P. Matikainen and M. Hebert., "Feature seeding for action recognition," in *ICCV*, 2011.

[10] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *CVPR*, 2008.

[11] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009.

[12] J. G. Liu, J. B. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild'," in *CVPR*, 2009.

[13] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," 2001.

[14] J. Zhang, M. Marszałek, M. Lazebnik, and C. Schmid, "Local features and kernel for classification of texture and object categories: A comprehensive study," vol. 73, pp. 213–238, 2007.

[15] L-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Objects as attributes for scene classification," in *ECCV Workshop Parts and Attributes*, 2010.